OzonMasters - Математическая статистика **(Due: 18/09/20)**

Лекция 1

Преподаватель: Владимир Панов *Автор конспекта:* Жибоедова Анастасия

1 Рекомендуемая литература

1. М.Б. Лагутин "Наглядная математическая статистика"
2. Robert Hogg, Elliot Tanis, Dale Zimmerman "Probability and Statistical Inference"
3. Yuri Suhov, Mark Kelbert "Basic Probability and Statistics"
4. Vladimir Spokoiny, Thorsten Dickhaus "Basics of Modern Mathematical Statistics"

2 Статистический эксперимент

2.1 Описание статистического эксперимента

Введем некоторые опредления для дальнейшего описания эксперимента с математической строгостью.

Определение 1. Ω - sample space, множество элементарных исходов.

Определение 2. $\mathcal{F}$ - σ-алгебры - множество подмножество пространства Ω, обладающее следующими свойствами:

1. $\emptyset, \Omega \in \mathcal{F}$
2. $A \in \mathcal{F} \Rightarrow \Omega \backslash A \in \mathcal{F}$
3. $A_1, A_2, \dots \in \mathcal{F} \Rightarrow \cup A_i, \cap A_i \in \mathcal{F}$

Определение 3. Борелевская σ-алгебры - это минимальное множество

Определение 4. Вероятностная мера $\mathbb{P} : \mathcal{F} -> [0, 1]$:

1. $\mathbb{P}\{\Omega\} = 1$,
2. $A_i, A_2 \dots \in \mathcal{F}, A_i \cap A_j = \emptyset \Rightarrow \mathbb{P}\{\cup A_i\} = \sum \mathbb{P}\{A_i\}$

Предположим, что вероятностная мера $\mathbb{P} \in \mathcal{P}$, если:

- $\mathcal{P} = \mathcal{P}_\theta = \{\mathbb{P}_\theta\}$ - модель параметрическая;
- $\mathcal{P}$ - пространство бесконечной размерности,$\mathcal{P} = \{$абсолютно непрерывная$\}$ - непараметрическая модель.

Определение 5. Статистические эксперимент - это модель, задаваемая следующей тройкой параметров: $(\Omega, \mathcal{F}, \mathcal{P})$. $\Omega, \mathcal{F}$ - описывают как проводится эксперимент, а $\mathcal{P}$ - оценивается по итогам эксперимента.

2.2 Примеры экспериментов

Реконструируем модели стандартных статистических экспериментов: подбрасывание монеты и выбор случайной точки на отрезке.

Эксперимент	Ω	$\mathcal{F}$	$\mathbb{P}$
Бросание монеты	$\{0, 1\}$	$\{\emptyset, \{0\}, \{1\}, \{0, 1\}\}$	$\mathbb{P}\{1\} = p; \mathbb{P}\{0\} = 1 - p$
Выбор точки на [0,1]	$[0, 1]$	$\{\{[a, b], 0 \leq a < b \leq 1\}, (a, b) = \Omega \backslash([0, a] \cup [b, 1]),$ [a,b) $= \cap(a - \frac{1}{n}, b), (a, b] = \cap(a, b + \frac{1}{n})$ {a} = [0,a] $\cap[a, 1]\}$ - Борелевская σ - алгебра	$\mathbb{P}\{[a, b]\} = b - a$ - нет косоглазия, иначе $\mathbb{P}$ - вер-ое распред. на [0,1]

3 Выборка (sample)

Определение 6. (*Простые учебники*) $x_1, \cdots x_n \in \mathbb{R}$ - выборка, это значения из набора случайных величин $X_1, \cdots X_n$ - i.i.d. (independent and identically distributed) с фиксированной $\omega \in \Omega$ (реализация случайных величин).

Примечание. На пространстве $(\Omega.\mathcal{F}, \mathbb{P})$ - нет независимых случайных величин, они либо тривиальны, либо зависимы. Чтобы провести статистический эксперимент, необходимо рассмотреть пространство $(\Omega \times \cdots \times \Omega), \mathcal{F} \times \cdots \times \mathcal{F}, \mathbb{P} : \mathbb{P}\{B \times \cdots \times B_n\} = \mathbb{P}_\theta(B_1) \cdot \cdots \cdot \mathbb{P}_\theta(B_n))$.

4 Описательная статистика (descriptive statistics)

Определение 7. $x_{(1)} \leq \cdots \leq x_{(n)}$, где $x_{(1)} = min(x_1, \cdots x_n)$ - варианционный ряд, порядковые статистики

Определение 8. Эсперическая p-квантиль - это такое число x, что:

- $\approx np$ чисел $< x$
- $\approx np$ чисел $> x$

Примечание. Понятие описательных статистик вышло за рамки описания параметров.

Определение 9. Медиана - (самый известный квантиль) $\frac{1}{2}$ - квантиль.

$$Med = \begin{cases} x_{(k)}, \text{если } n = 2k+1 \\ \dfrac{x_{(k)} + x_{(k+1)}}{2}, \text{если } n = 2k \end{cases}$$

4.1 Оценивание квантилей

offtop

Абсолютно непрерывные распредления, это такие распределения, что функция распредления имеет производную $F'(x) = p(x)$, где $p(x)$ - плотность распредления (probability density function).

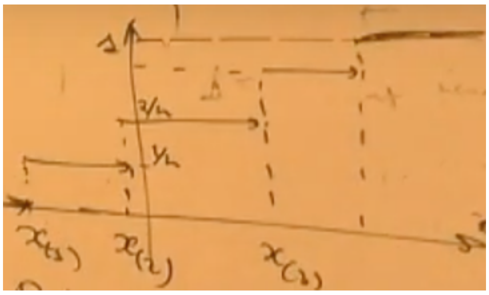
Собирается статистика количества клиентов в магазине без понимания вида распределения. В случае, если распределение никаким образом не ограничивается параметрами, оно считается непараметрическим.

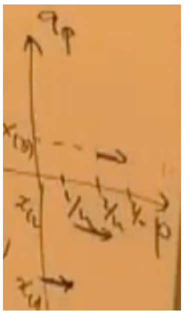
Из определения квантиля можно увидеть то, что нам необходимо решить уравнение $F(x) = p$ (cumulative distribution function) - , где $\mathbb{P}\{X \leq x\}$. Решение уравнения возможно двумя способами: из книжек и из пакетов.

4.1.1 Способ 1

Теоретический способ решения уравнения. $\hat{F}_n(x)$ - empirical distribution function - оценка функции $F(x)$.

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{X_i \leq x\}$$





Решить уравнение из предположения $\hat{F}_n(x) = F(x)$, тогда $\hat{F}_n(x) = p$:

$$x = q_p = \begin{cases} x_{(pn)}, pn \in \mathbb{N} \\ x_{(\lfloor pn \rfloor + 1)}, pn \notin \mathbb{N} \end{cases}$$

Иногда встречается упрощенная модификация решения (выборочная α - квантиль) $q_p = x_{(\lfloor pn \rfloor + 1)}$

Недостатком данного способа решения является то, что на выходе получается разрывная функция q_p.

4.1.2 Способ 2

Определим функцию следующим образом:

- определим значения функций в наборе точек как - $(, x_{(k)}), k = 1, \cdots n$;
- доопределим функцию на отрезках линейными функциями.

В описанном подходе точки выбраны из значений функции эмперического распредления, но в пакетах используются иные значения.

Построим последовательность $F(x_{(1)}), F(x_{(2)}) \cdots F(x_{(n)})$ - n чисел, являющихся реализациями случайных величин $F(X_{(1)}), F(X_{(2)}) \cdots F(X_{(n)})$. Поскольку функция $F(X_{(i)})$ - монотонно возрастающая, то можно трактовать это так, что послежовательность i.i.d случайных величин $F(X_{(1)}), F(X_{(2)}) \cdots F(X_{(n)})$ соответствует расположению в порядке возрастания случайных величин $F(X_1), F(X_2) \cdots F(X_n)$. $\mathbb{P}\{F(X) \le x\} = \mathbb{P}\{X \le F^{-1}(x)\} = F(F^{-1}(x)) = x$, тогда с.в. $F(X_1), F(X_2) \cdots F(X_n)$ - имеют равномерные распределения.

Построим следующую цепочку выводов (без доказательств):

- $U_1 \cdots U_n - Unif([0, 1])$ равномерно распределенные с.в.,
- тогда $U_{(i)} \sim Beta(i, n - i + 1)$ - i-ый член вариационного ряда
- тогда $p_i(x) = \dfrac{n!}{(i-1)!(n-i)!} x^{i-1}(1-x)^{n-i}$ плотность случайной величины $F(X_{(i)})$
- можно доказать, что $\mathbb{E}U_i = \mathbb{E}F(X_{(i)} = \dfrac{i}{n+1}$
- из вышесказанного следует, что в качестве опорных точек построения квантилей можно взять $(\frac{k}{n+1}, x_{(k)}), k = 1, \cdots n$ - реализация пакета spss (и др.).

Почему характерным значением называется среднее??? Оказывается, что это не совсем так.

Определение 10. Мода абсолютно непрерывного распределения - точка максимума плотности (наиболее модное значение).

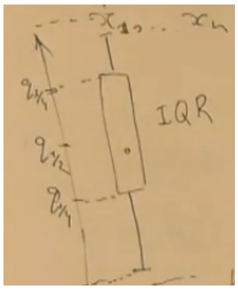
Во многих пакетах вместо среднего берется мода - $mode(U_i) = \frac{i-1}{n-1}$, соответственно в качестве опорных точек $(\frac{k-1}{n-1}, x_{(k)}), k = 1, \cdots n$.

5 BoxPlot - box and whiskers

Определение 11. IQR (interval quartile range) - межквартильный размах, расстояние между $\dfrac{1}{4}$ - квантилем (нижний квартиль) и $\dfrac{3}{4}$ - квантилем (верхний квартиль).

BoxPlot строится для выборки $x_1, \cdots x_n$ следующим образом:

1. коробка - на графике функции отмечается $\dfrac{1}{2}$ - квантиль и вокруг него строится межквартильный размах (IQR) в виде прямоугольника;
2. усы - отменяается ближайшее сверху значение из выборки к границе $q_{\frac{1}{4}} - 1.5 * IQR$ и ближайшее снизу значение из выборки к границе $q_{\frac{3}{4}} + 1.5 * IQR$, эти значения являются концами усов



Определение 12. Значения, не попавшие в boxplot называются **выбросами**.

Примечание. Чем меньше коробка, тем больше мы можем доверять медиане. Рассмотрим на примере подбрасывания монеты. Собирается статистика эксперимента. Полученная медиана выборки = 30 (30 раз выпал орел), но выборка имеет очень большой разброс (IQR - большой, относительно медианы). Это означает, что значению медиане нельзя достаточно доверять и результаты эксперимента не надежны.

6 Оценивание параметров

6.1 Постановка задачи

Дан набор $X_1, \cdots X_n$ - i.i.d., случайные величины имеют параметрическое распределение $\sim \mathbb{P}_{\overrightarrow{\theta}}$. Рассматривая выборку $x_1, x_2 \cdots x_n$ - оценить параметр $\overrightarrow{\theta}$.

6.2 Метод моментов

Рассматриваем математическое ожидание некоторой функции - $\mathbb{E}g(X) = m(\overrightarrow{\theta})$, где $X \sim X_i$. Оцениваем математическое ожидание как: $\frac{1}{n}\sum g(x_i) = m(\hat{\theta}_n)$, тогда $\hat{\theta}_n = m^{(-1)}(\frac{1}{n}\sum g(x_i))$

Пример 1

Задано нормальное распредление $\mathcal{N}(\mu, 1)$. Предположим $g(x) = x$, тогда $\mathbb{E}(X) = \mu \Rightarrow \frac{1}{n}\sum x_i = \hat{\mu}_n$.

Пример 2

Задано нормальное распредление $\mathcal{N}(\mu, \sigma^2)$. Предположим $g(x) = x$, тогда $\mathbb{E}(X) = \mu \Rightarrow \frac{1}{n}\sum x_i = \hat{\mu}_n$ и $g(x) = x^2$, тогда $\mathbb{E}(X^2) = \mu^2 + \sigma^2 \Rightarrow \frac{1}{n}\sum x_i, 2 = \mu_n^2 + \sigma_n^2 \Rightarrow \hat{\sigma}_n^2 = \frac{1}{n}\sum x_i^2 - \hat{\mu}_n^2$

6.3 Метод максимального правдоподобия

Определение 13. $L(\overrightarrow{\theta}) = \Pi p_\theta(x_i)$ - функция правдоподобия.

Метод заключается в формировании предположение о плотности распредления набора i.i.d.$(X_1, \cdots X_n \sim p(\overrightarrow{x})$.

Пример 1

Задано нормальное распредление $\mathcal{N}(\mu, 1)$. Предположим $p_\mu(x) = \dfrac{1}{\sqrt{2\pi\sigma}} \exp -\frac{(x-\mu)^2}{2\sigma^2}$, тогда

$$\log L(\overrightarrow{\theta}) = \sum \log\left(\frac{1}{\sqrt{2\pi\sigma}} \exp -\frac{(x-\mu)^2}{2\sigma^2}\right) = n\log\left(\frac{1}{\sqrt{2\pi\sigma}}\right) - \sum\frac{(x-\mu)^2}{2\sigma^2} \to max_\mu$$

Примечание. Пакеты работают по следующей логике: на вход подается подсчитанная функция и внутри пакета рассчитывается ее максимум или минимум по параметру.

6.4 Экспоненциальное семейство распредлений

Определение 14. Семейство распределений $\mathcal{P} - (p_v(x))$ - называется экспоненциальным, если $\exists g, d$ - функции, такие что $p_v(x) = g(x)e^{xv - d(v)}$

Пример 1

Задано нормальное распредление $\mathcal{N}(\mu, 1)$. Предположим

$$p_\mu(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2}\right) = \frac{1}{\sqrt{2\pi\sigma}}\exp\left(-\frac{-x + 2x\mu - \mu^2}{2}\right) = \frac{1}{\sqrt{2\pi\sigma}}\exp\left(\frac{-x}{2}\right)\exp\left(x\mu - \frac{\mu^2}{2}\right)$$

Тогда $v = \mu, d(v) = \dfrac{\mu^2}{2} = \dfrac{v^2}{2}$

Пример 2

Схема Бернулли (подрасывание монеты) - дискретное распредление $p_v(x) = \mathbb{P}X = x$.

$$p_v(x) = \begin{cases} p, x = 1 \\ 1 - p, x = 0 \end{cases} = p^x(1-p)^{1-x} = e^{x\log p}e^{(1-x)\log(1-p)} = e^{x\log\frac{p}{1-p}}e^{\log(1-p)}$$

Обозначим $v = \dfrac{p}{1-p} \Rightarrow p = \dfrac{e^v}{e^v + 1}$, тогда $d(v) = -log(1-p) = -\log\left(1 - \dfrac{e^v}{e^v+1}\right) = \log(e^v + 1)$

Утверждение 1. $d'(v) = \mathbb{E}(\xi), \xi \sim p_v(x)$, $d''(v) = \mathbb{D}(\xi)$.

Доказательство. —-

Из первого примера $d(v) = \frac{v^2}{2} \Rightarrow d'(v) = v = \mu, d''(v) = 1$.

Из второго примера $d(v) = \log(e^v + 1) \Rightarrow d'(v) = \frac{e^v}{e^v + 1} = p$, $d''(v) = \frac{e^v}{(e^v + 1)^2} = p(1-p)$

6.4.1 Метод моментов для э.с.р.

$\mathbb{E}\xi = d'(v)$, тогда $\frac{1}{n}\sum x_i = d'(\hat{v}_n)$.

Поскольку $d''(v) \leq 0$ (дисперсия), тогда $d'(v)$ - монотонно возрастающая и непрерывная (поскольку имеет производную) $\Rightarrow \exists (d')^(-1) \Rightarrow \hat{v_n} = (d')^{(-1)}\left(\frac{1}{n}\sum x_i\right)$

6.4.2 Метод максимального правдоподобия для э.с.р.

$L(\hat{v}) = \Pi p_v(x_i) = \frac{1}{n}\sum \log(g(x_i)e^{x_i v - d(v)}) = \frac{1}{n}\sum \log(g(x_i)) + \frac{1}{n}\sum(x_i v - d(v))$

$\log L(v) = argmax\frac{1}{n}\sum(x_i v - d(v))$, обозначим $G(v) = \frac{1}{n}\sum(x_i v - d(v))$, тогда $G'(v) = \frac{1}{n}\sum x_i - d'(\hat{v_n}) = 0$, $\hat{v_n} = (d')^{(-1)}\left(\frac{1}{n}\sum x_i\right)$

7 Достаточная статистика

Может ли функция от выборки заменить саму выборку? Что если нам известны только агрегированные параметры (метапараметры)?

Определение 15. Достаточная статистика (sufficient statistics) - статистика $T(x_1, \cdots x_n)$ называется достаточной, если $\mathbb{P}\{(X_1, \cdots X_n) \in B | T(X_1, \cdots X_n) = t\}$ не зависит от параметров распредления $B \in \mathbb{R}^n$

. ***Критерий факторизации***. *Статистика называется достаточной тогда и только тогда, когда* $p_v(x_1) \cdot \cdots p_v(x_n) = g(T(x_1, \cdots x_n), \theta)h(x_1, \cdots x_n)$

Выборка моделируется следующим образом:

1. выбирается значением статистики $T(X_1, \cdots X_n)$.
2. зная условное распредление строю выборку на основе информации о статистике.

Пример 1

Найдем достаточную статистику для экспоненциального семейства распредления, если она существует. Запишем произведение плотностей из критерия факторизации:

$$p_v(x_1) * \cdots p_v(x_n) = g(x_1) * \cdots g(x_n)e^{v\sum x_i - nd(v)}$$

Тогда видим, что функция зависит неподредственно от суммы выборки и не зависит от каждого значения в отдельности. По критерию факторизации достаточная статистика $T = \sum x_i$

Пример 2

Выборка из ранвомерного распредления на отрезке $[0, \theta]$, $X_1, \cdots X_n \sim Unif([0, \theta])$. Тогда $p_\theta(x) = \frac{1}{\theta}\mathbb{I}\{x \in [0, \theta]\}$. Применим критерий факторизации:

$$p_v(x_1) \cdots p_v(x_n) = \frac{1}{\theta^n}\mathbb{I}\{x_1 \in [0, \theta] \cdots x_n \in [0, \theta]\} = \frac{1}{\theta^n}\mathbb{I}\{max(x_1, \cdots x_n) < \theta\}$$

. Получаем, что $T = max(x_1, \cdots x_n)$ - достаточная статистика.